

Predictive Data Management (PDM) makes profiling and data testing more simple, powerful, and cost effective than ever before. Version 6.0.5 adds new SOA and in-stream capabilities while delivering a powerful new viewer simplifying its use. PDM is designed for data stewards, data architects, business analysts, consultants and any individual challenged with the need to analyze source data for anomalies. This discovery work is critical to the success of projects that require data migration, such as ERP, CRM, EDI, SCM, data warehousing / business intelligence, MDM, and data governance.

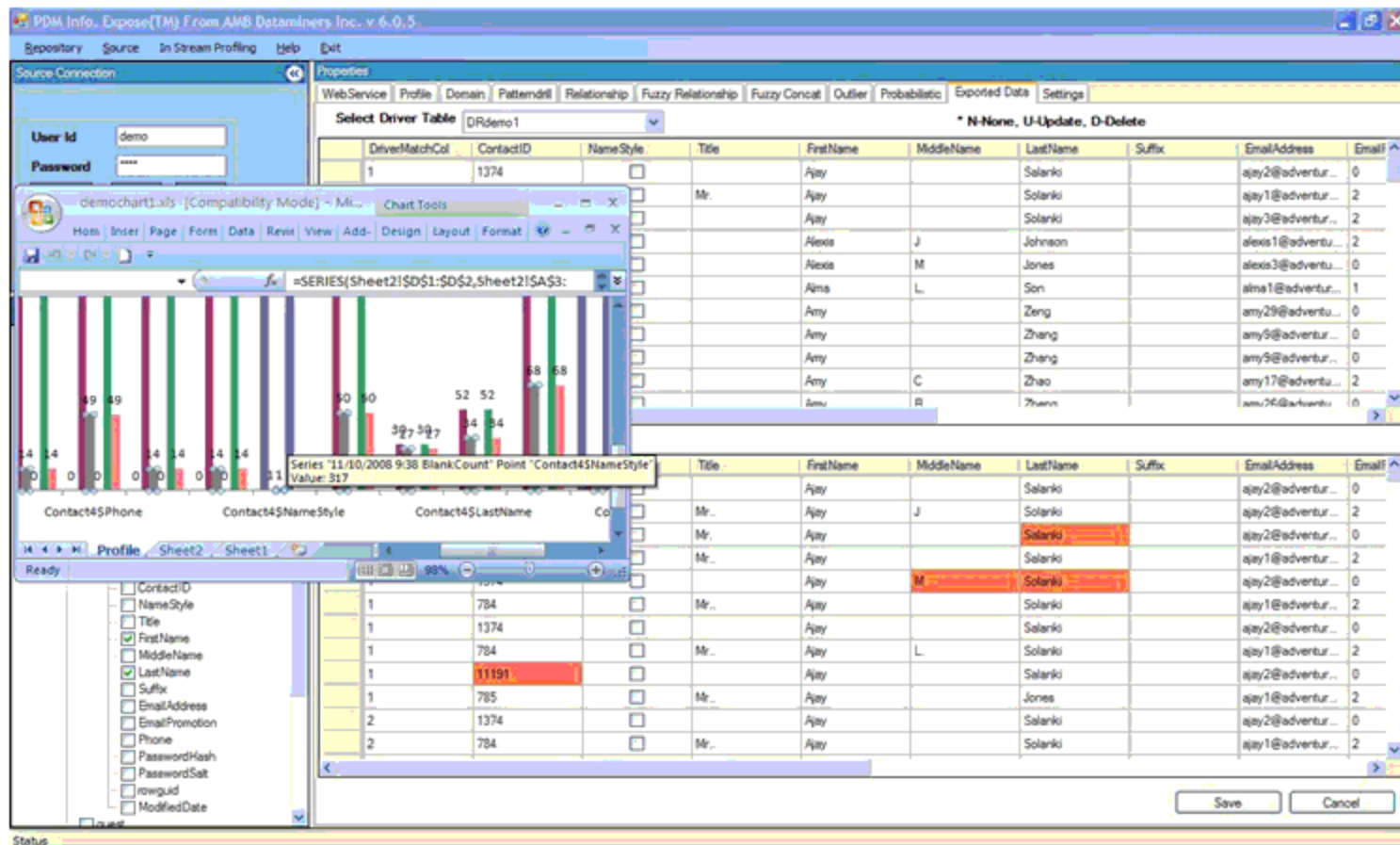
All Profiling options: Web Services (SOA), batch, in-stream, within-Excel, and within-ETL:

PDM Version 6.0.5 now offers all profiling options: SOA, batch, in-stream, and within-Excel profiling/data cleansing. All functions that are part of PDM are supported both in batch and with a SOA Web Services layer. Through SOA, users create and view profiling statistics dynamically in real time. Additionally, in-house programs are able to call the SOA and execute all PDM functions of profiling and analysis. This ensures data governance can be carried out throughout your organization, both operational and analytic.

InfoExpose (PDM's Business Intelligence Viewer and Drill-back Analysis Tool):

Users easily view and analyze profile results and the underlying data with InfoExpose. Non-technical individuals, subject matter experts and casual users can become sophisticated information consumers. InfoExpose displays data anomalies and supports drill back to the actual source records to determine the best course of action. Users can easily export InfoExpose information into Excel, Pivot tables, or any desktop tool of choice.

Users Export InfoExpose Results to Excel and Pivot Tables.



Profiling/Data Analysis directly at the Source and stored in an Open non proprietary Repository
PDM stores three categories of statistics in an open non-proprietary metadata database.

1. Technical Metadata (i.e. table size, number of columns, column size, key info)
2. Column information (i.e. min/max, duplicate counts, blanks, special characters, patterns, unique domain counts, unique patterns count)
3. User Statistics (user defined business rules and statistics).

InfoExpose Shows Profiling Results, Drills Back on Anomalies and Does Graphical Reporting

The screenshot displays the PDM Info. Expose(TM) application window. The main area shows a table of column statistics for a table named 'Contact4'. The table has columns for Column, Maximum Value, Minimum Value, RowCnt, BlankCount, Duplicate, Unique, Zero, HasSpecialChar, UniqueDom, PercentUnique, DomainCount, MeanValue, and StandardDeviation. The 'Duplicate' column for 'LastName' is highlighted in blue, indicating a value of 34.

Below the table is a bar chart showing the distribution of values for various columns. The x-axis lists columns like ContactID, NameStyle, Title, etc., and the y-axis shows counts. The chart is titled 'DuplicateCount for [dbo].[Contact4].[LastName] in SQL SERVER Data...'. A small window is open over the chart, showing a detailed view of the duplicate counts for the 'LastName' field.

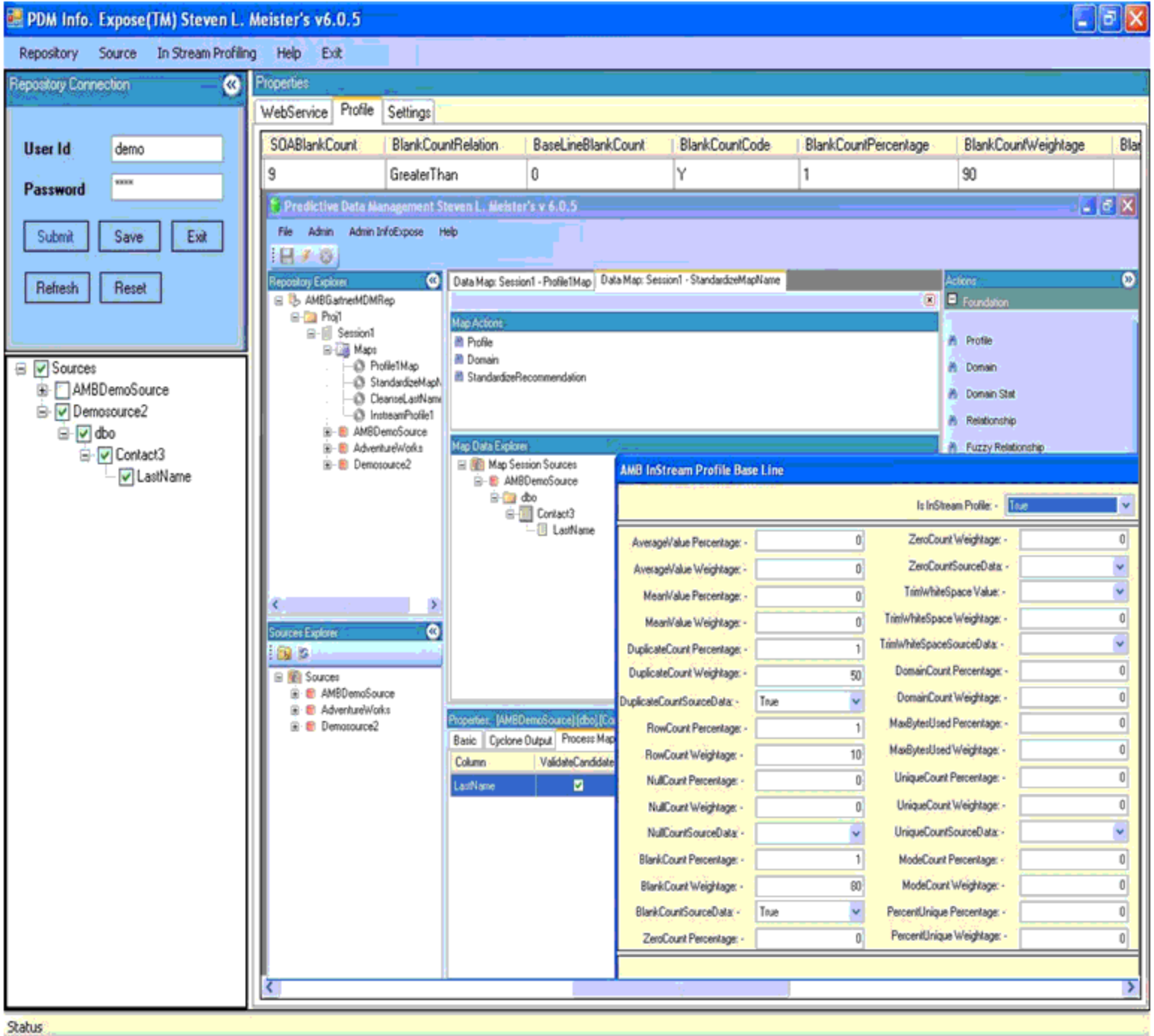
Column	Maximum Value	Minimum Value	RowCnt	BlankCount	Duplicate	Unique	Zero	HasSpecialChar	UniqueDom	PercentUnique	DomainCount	MeanValue	StandardDeviation
ContactID	19989	1	348	0	22	257	0	<input type="checkbox"/>	5	74	279	11604.37356321...	7557.9435
NameStyle	0	0	348	348	1	0	348	<input type="checkbox"/>	1	0	1	0	0
Title	Sra.		348	16	4	1	0	<input checked="" type="checkbox"/>	4	0	5	0	0
FirstName	Zachary	Abigail	348	0	68	99	0	<input type="checkbox"/>	8	28	167	0	0
MiddleName	W	A	348	0	27	3	0	<input checked="" type="checkbox"/>	4	1	30	0	0
LastName	Zhu	Smith	348	0	34	18	0	<input checked="" type="checkbox"/>	14	5	52	0	0
Suffix			348	0	0	0	0	<input type="checkbox"/>	1	0	0	0	0

FirstName	MiddleName	LastName	Suffix	EmailAddress
Kim		Abercrombie		kim2@adventure...
Kim	B	Abercrombie		kim1@adventure...
Kim		Abercrombie		kim7@adventure...
Hazem	E	Abolrous		hazem0@advent...
Sam		Abolrous		sam1@adventur...
Pilar		Ackeman		pilar1@adventur...
Pilar	G	Ackeman		pilar0@adventur...
Jay	G	Adams		jay0@adventure...

AMB-SOA offers Ongoing Data Governance with In-Stream Profiling:

In-stream captures dynamic profiling results and compares them to pre-defined baseline standards pre-set by the user. By using a combination of a weighting methodology and baseline standards and tolerances, PDM can automate the decision making as to how anomalies should be treated in real-time processes, while protecting target applications from non-compliant data.

SOA In-Stream Establishes Baseline Standards to Evaluate Data Anomalies In-stream



Outlier Discovery in InfoExpose / SOA – Identify unexpected or out of range values

Dynamic Outlier Discovery is a unique feature of InfoExpose. Within the profiled statistics, PDM will assist users by detecting outliers potentially skewing application results. Outliers are determined by the standard deviation of each column and the calculated sigma for each unique value. The user then established rules on determining what is a “true” outlier for their unique data process.

The screenshot displays the AMB-PDM software interface with three overlapping windows. The top window shows a table of statistics for the 'SalesOrderHeader' table, specifically for the 'TotalDue' column. The bottom window shows a list of outliers for the same column, with a 'Value' of 14987.07269679909 highlighted. The middle window shows a list of columns in the 'SalesOrderHeader' table, with 'TotalDue' selected.

TableName	ColumnName	DuplicateCount	MinimumValue	MaximumValue	RowCount	StandardDeviation	MeanValue	Percentage
SalesOrderHeader	TotalDue	1189	1.5183	247913.9138	31465	14987.07269679...	4471.876206089...	11

DatabaseName	TableName	ColumnName	RunTime	Selected Value	Operator	Description	Percentage
AdventureWorks	SalesOrderHeader	TotalDue	Percent	3	>=	pctsignmag3	3.08914667090417924678...
AdventureWorks	SalesOrderHeader	TotalDue	Value Amount	3	>=	amountsignmag3	53.44

TableName	ColumnName	Frequency	Value	Value_Length	Pattern	DiscretePCT	DomainSigma	TimeRun
SalesOrderHeader	TotalDue	1	247913.9138	9	999999.9999	0.11	15.54185035433...	11/27/2008
SalesOrderHeader	TotalDue	1	227737.7215	9	999999.9999	0.11	15.19561065107...	11/27/2008

ShipMethodID	CreditCardID	CreditCardApproval	CurrencyRateID	SubTotal	TaxAmt	Freight	TotalDue
5	4830	27578V25371		224356.4831	17948.5186	5608.9121	247913.9138

TableName	ColumnName	Frequency	Value	Value_Length	Pattern	DiscretePCT	DomainSigma	TimeRun
SalesOrderHeader	TotalDue	1	157212.8539	9	999999.9999	0.11	10.48989733222...	11/27/2008
SalesOrderHeader	TotalDue	1	156203.7767	9	999999.9999	0.11	10.422567493267...	11/27/2008
SalesOrderHeader	TotalDue	1	155260.0958	9	999999.9999	0.11	10.35960116702...	11/27/2008
SalesOrderHeader	TotalDue	1	154912.0712	9	999999.9999	0.11	10.33637951413...	11/27/2008
SalesOrderHeader	TotalDue	1	153845.2385	9	999999.9999	0.11	10.26519578695...	11/27/2008
SalesOrderHeader	TotalDue	1	153095.214	9	999999.9999	0.11	10.21248225697...	11/27/2008
SalesOrderHeader	TotalDue	1	150167.6656	9	999999.9999	0.11	10.101981299737...	11/27/2008
SalesOrderHeader	TotalDue	1	149897.3647	9	999999.9999	0.11	10.00177738058...	11/27/2008
SalesOrderHeader	TotalDue	1	149061.0659	9	999999.9999	0.11	9.999355373247...	11/27/2008
SalesOrderHeader	TotalDue	1	148463.7645	9	999999.9999	0.11	9.972845766732...	11/27/2008
SalesOrderHeader	TotalDue	1	148423.9563	9	999999.9999	0.11	9.970188687592...	11/27/2008

Excel Add-In allows profiling directly within Excel spreadsheets:

Most profiling today is accomplished through Excel spreadsheets at some step. PDM's Excel Add-In profiles and analyzes data right from within Excel worksheets or data sources and puts results into a new profiling worksheet.

The screenshot shows two overlapping Excel windows. The top window displays the 'customer_dim' Add-In menu with options like 'Cube Analysis', 'Start Meeting Now', and 'Add Source'. The bottom window shows a data table with columns for 'customer_key', 'customer_name', 'customer_contact', etc., and a corresponding 'AMB Profile Result' sheet.

ColumnName	UniqueCount	AverageValue	BlankCount	DomainCount	DuplicateCount	TrimWhiteSpace	MinimumValue	MaximumValue	MaxByte
customer_key	18	10	0	18	0	FALSE	1	18	
customer_nbr	18	285623	0	18	0	FALSE	171684	502301	
customer_name	18	18	0	18	0	FALSE	April's Showers	Yoyodyne Industries	
customer_contact	18	12	0	18	0	FALSE	A. Bomb	Vincent Price	

customer_key	customer_nbr	customer_name	customer_contact	city	state	postal_code	country	phone	fax	default_province	
1	171684	Nostrum Mr. Van N	CEO	Stratham	NH	3885		6.04E+09	6.04E+09	6 NE	
2	188768	Hellow Kri Bill	President	Phhhbttt	16833 ANTWERP RO	OCEANSIC	CA	8.01E+09	7.61E+09	60 2%	
3	180226	Vandelay Art Vande	CEO	#611-525	3000 CABOT BLVD W	PORTLAN	OR	5.03E+09	5.03E+09	33 1%	
4	184497	Kramerica Kramer	President	16727 SW	3500 N O'HENRY BLV	ALBUQUE	NM	8.01E+09	8E+09	4 NE	
5	173955	Yoyodyne John Who	President	123 fake street	Grovers M	NJ		5.04E+09	5.04E+09	10 NE	
6	193039	Coffee Be Brad Ema	Minding N	239 UTAH	8520 PAN AM FRWY	LAKE ZURI	IL	8.48E+09	8.48E+09	104 ON	
7	197310	Nukes-r-U.A. Bomb	Call securi	P.O. BOX	(20209 BROADWAY	RYE	NY	9.15E+09	9.15E+09	104 ON	
8	201581	Wax Work Vincent P	Grrrrrr	4820 BLAL	P.O. BOX 915	OGDENSB	NY	8.2E+09	8.77E+09	6 NE	
9	205852	Fzzbrbl Kristin Pz	dpelski	P O BOX	509	KEESEVILL	NY	8E+09	5.19E+09	32 1.5'	
10	210123	April's Shc April How	ell	201 WEST	P.O. BOX ;32057 64T	ST. LOUIS	MO	63103		17 1%	
11	214394	Cthulu Co David B Ev	Shining Il	1466 BROADWAY	STE. 800	CHICAGO	IL	60631		79 NE	
12	406995	Shoggoth Michelle	PRESIDEN	101 INDUSTRIAL	ST	PITTSFIE	MA	1201	4.13E+09	4.13E+09	6 NE
13	412502	Deep One Monica	Eye of the	PO BOX	367	HOLLYWO	FL	33020	8.01E+09	9.55E+09	68 LTR
14	417101	Max Burtc Barbara /	ennifer	GINA SUN	3274 BEEKMAN ST	MIAMI	FL	33138	3.06E+09	3.06E+09	6 NE
15	425590	Wigwam Mrs Alber	CEO	8786 WAT	8786 WATER ST	ASHLAND	OR	97520	5.03E+09		104 ON
16	426402	New Zeal Cam Nich	SHIPPING	18245	BUTTEVILLE ROAD NE	MEDIA	PA	19063	6.11E+09	6.11E+09	79 NE
17	426902	Paw Paw Michelle	VICE PRES	4260	JUSTIN ROAD	EDEN PRA	MN	55346	6.13E+09	6.13E+09	79 NE
18	502301	Luxe Desi Ben Van	GENERAL	P.O. BOX	407	LANGHOR	PA	19047	2.16E+09	2.16E+09	6 NE
19	502900	Vintage Bi Butch		Hey You!		SEATTLE	WA	98101	2.07E+09	2.07E+09	6 NE
20	516920	Telluride Brian Freon		1730 W	WRIGHTWOOD	SAN FRAN	CA	94103	4.15E+09	4.15E+09	7 NE
21	601098	J D Intern Bob Howa	OWNER	357	SUTTER AVE	SOLANA B	CA	92075	6.19E+09	6.19E+09	79 NE
22	605401	Sandra Mi Rodney Ki	PRESIDEN	375	ALAB; 22926 CIELO VISTA D	SANDY	UT	84070	5.03E+09	5.03E+09	104 ON
23	608101	Stansbury Samantha	VP SALES	Hey You!		BERGENFI	NJ	7621	2.01E+09	2.01E+09	67 1%
24	618402	Gki / Beth Bernie No	SALES MA	690	N. 5TH ST.	NEW YORK	NY	10173	2.13E+09	2.13E+09	54 2%
25	623099	Dimensioi N. Dimen	SALES MA	P.O. BOX	227	VERNON	CA	90058	2.14E+09	2.14E+09	6 NE
26	624102	Ore-iginal Rose Pero	Former pr	9467	NE VIEWCREST	CARROLT	TX	75006	8.01E+09	8E+09	7 NE

The MDM Discovery Solution: Match and merge similar or potential duplicate values:

Classic profiling is only the first step in developing master data management (MDM) processes. PDM can match and maintain similar value(s) across multiple tables. Users view tables side by side and use drag and drop techniques to perform maintenance functions. This stores and builds business rules in cases where you are merging multiple master file data, MDM/ETL processes, or any Data Management activity.

InfoExpose Display of Match/Merge - Graphical View and Associated Business Rules

The screenshot displays the AMB-PDM software interface. At the top, the title bar reads "PDM Info. Expose(TM) Steven L. Meister's v6.0.5". The main window is divided into several panes:

- Source Connection:** Fields for User Id (demo) and Password, with buttons for Submit, Save, Next, Refresh, Reset, and Exit.
- Source DB Explorer:** A tree view showing a "Contact" table with fields like ContactID, NameStyle, Title, FirstName, MiddleName, LastName, and Suffix.
- Properties:** A tabbed interface with "Probabilistic" selected. It shows "Select Driver Table" as DRdemo1 and a list of columns: DriverMatchCol, ContactID, NameStyle, Title, FirstName, MiddleName, LastName, Suffix.
- Match Table Name:** Set to MRdemo1.
- Match Table:** A table showing 9 rows of data with columns: DriverMatchCol, ContactID, NameStyle, Title, FirstName, MiddleName, LastName, Suffix.
- Match Explorer Details:** A table listing match results with columns: DriverTable, DriverCol, DriverValue, MatchId, MatchTable, MatchCol, MatchOriginalValue, MatchModifiedValue, MatchPercent, User Id, User Name, TimeRun.
- Match List:** A list of matches for the "LastName" column, showing "Solanki" and "Zhang" with a "Cancel" button.

Find duplicate and similar data through fuzzy probabilistic matching in InfoExpose:

InfoExpose uses an advanced Fuzzy Algorithm to assist in finding duplicated data. By stating and requesting a specific probability, PDM finds all values that meet or exceed that percentage and will drill back to the source data to validate the results. This is key for organizations consolidating master file data like customer or item files where analysis across various tables is required. This becomes a repeatable process as PDM develops the business rules for future consolidation. See also the description of visual match merge on the previous page.

InfoExpose Display of Probabilistic Fuzzy Match - Graphical View and Source Drill Back

The screenshot displays the PDM Info. Expose v6.0.5 application window. The main interface is divided into several panes:

- Source Connection:** Shows 'User Id' as 'demo' and 'Password' as '****'. Buttons for 'Submit', 'Save', 'Refresh', and 'Reset' are visible.
- Properties:** A tabbed interface with 'Probabilistic' selected. It includes export options like 'ExportToExcel', 'ExportToCSV', 'ExportToHTML', 'ExportToTXT', 'ExportToXML', and 'Export Merge'.
- Main Table:** A table with columns: DriverTable, DriverColumnName, DriverValue, MatchTable, MatchValue, NumberMatches, MatchingPercent, and TimeRun. It lists several matches for 'Contact3' with 'FirstNameLastName' values like 'AjaySalanki' and 'AlmaSon'.
- Source DB Explorer:** A tree view on the left showing a 'Contact' table with various fields like 'ContactID', 'NameStyle', 'Title', 'FirstName', etc., checked for display.
- MatchValue for [dbo].[Contact].[FirstName_LastName] in [Sqlserver]:** A drill-back window showing a table with columns: ContactID, Title, FirstName, MiddleName, LastName, EmailAddress, Phone, and PasswordHash. It lists records for 'Ajay Solanki' with ContactIDs 784 and 1483.

Identify Profiling/Quality anomalies directly within the SSIS Pipeline Architecture:

PDM provides integration with Microsoft SQL Server Integration Services (SSIS). Developers easily drag and drop profiling/quality actions as part of an SSIS data flow directly into the SSIS environment—thus directly testing data transformation during development. PDM SOA Web Services as well as data cleansing are fully imbedded in the SSIS toolbox and are part of the SSIS pipeline architecture.

SSIS display of the AMB-PDM Profiling Data Flow with Graphical Interface and Output Results

The screenshot shows the SSIS Data Flow Task in Visual Studio. The Data Flow Task includes a DataReader, AMB WebServices, and a Destination. The Properties window for the AMB WebServices component is visible, showing the following details:

Table Name	Column Name	Unique Count	Average Value	Domain Count	Duplicate	Trim White Space	Minimum Value	Maximum Value	Standard Deviation	Mean Value
ContactID	ContactID	54	11863	66	12	<input type="checkbox"/>	1	19962	8910.519023885...	11863.345794
ContactID	EmailAddress	54	0	66	12	<input type="checkbox"/>	123456789	william5@adventure-wor...	0	0
ContactID	FirstName	39	0	54	15	<input type="checkbox"/>	Abigail	William	0	0
ContactID	LastName	32	0	49	17	<input checked="" type="checkbox"/>	Smith	Zhou	0	0

The screenshot shows the Data Viewer for the AMB WebServices component. The Data Viewer displays a table with the following columns: ColumnName, UniqueCount, DomainCount, DuplicateCount, TrimWhiteSpace, MinimumValue, and MaximumValue. The table contains 4 rows of data:

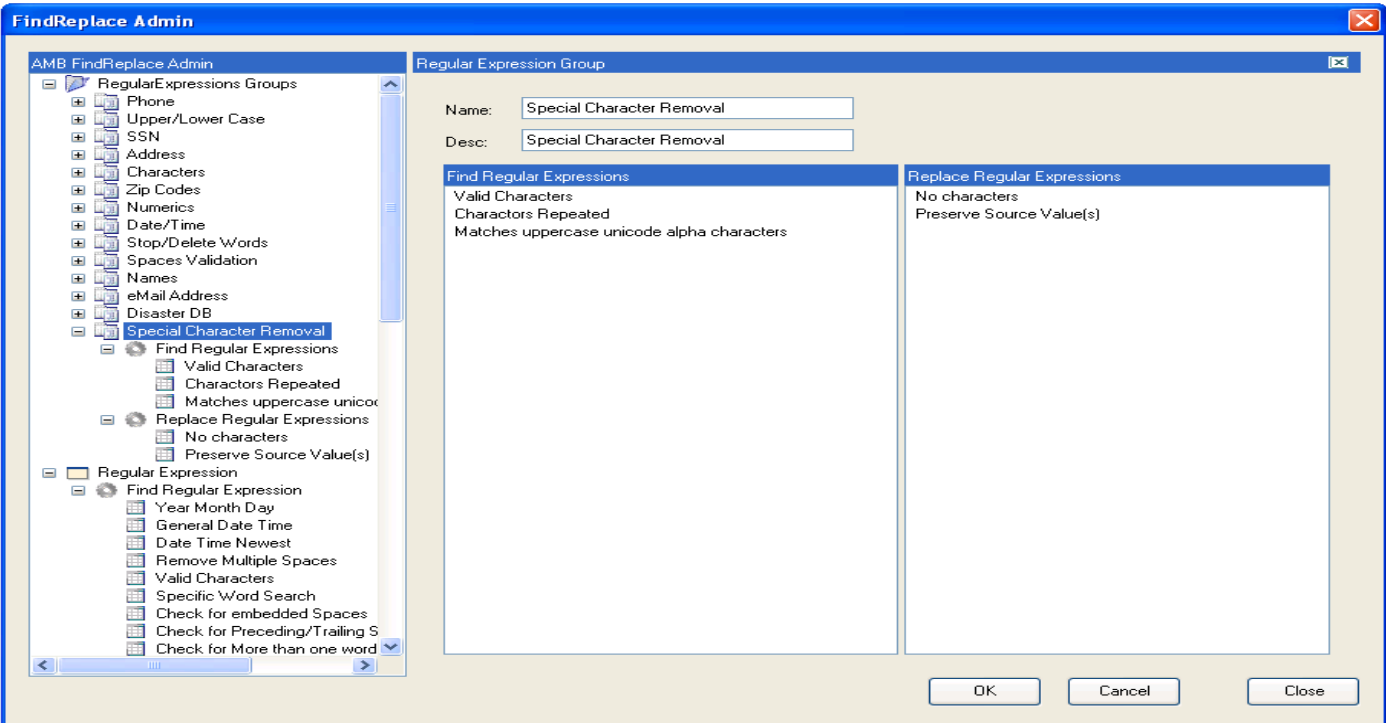
ColumnName	UniqueCount	DomainCount	DuplicateCount	TrimWhiteSpace	MinimumValue	MaximumValue
ContactID	54	66	12	False	1	19962
EmailAddress	54	66	12	False	123456789	william5@adventure-wor...
FirstName	39	54	15	False	Abigail	William
LastName	32	49	17	True	Smith	Zhou

Other Significant Features

Business Rules Automation for Source Data Validation:

PDM uses a Regular Expressions Engine to ensure that your source data is properly validated /formatted. PDM is packaged with a set of regular expressions, and also permits users to incorporate their own regular expressions. A good example of regular expressions is zip code format.

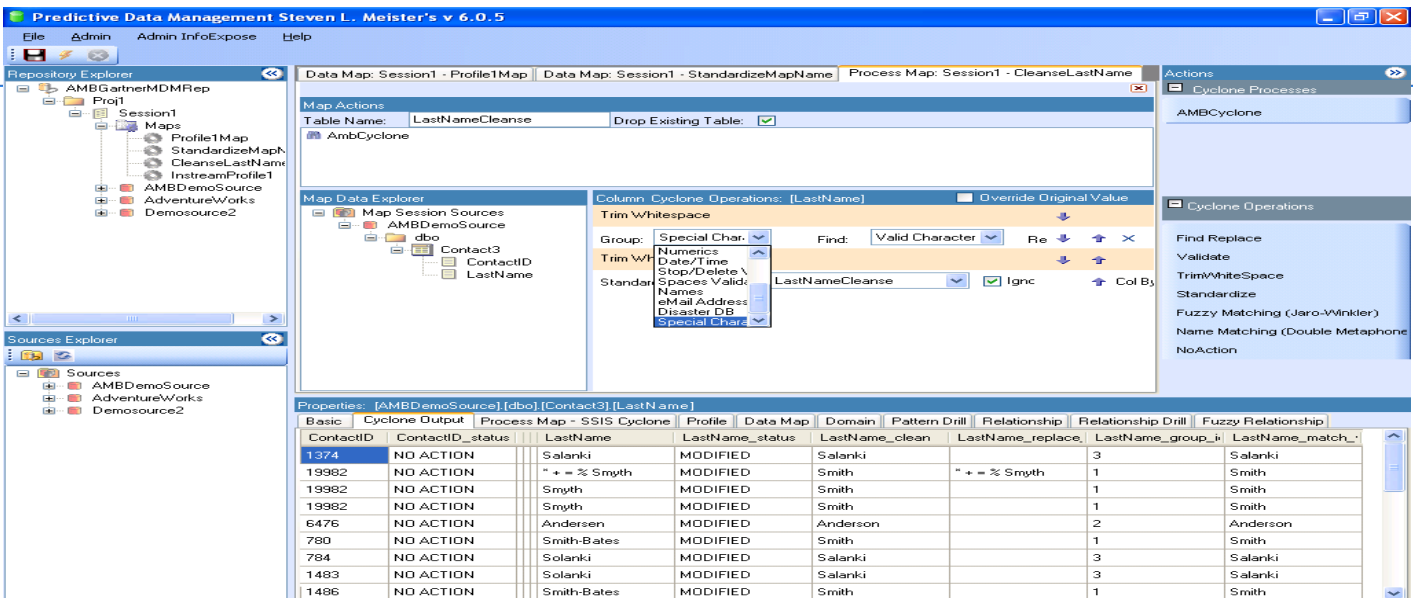
PDM Regular Expression Engine with User Defined Business Rules Results



The screenshot shows the 'FindReplace Admin' dialog box. On the left is a tree view of 'Regular Expressions Groups' with 'Special Character Removal' selected. The main area is titled 'Regular Expression Group' and contains:

- Name: Special Character Removal
- Desc: Special Character Removal
- Find Regular Expressions:
 - Valid Characters
 - Characters Repeated
 - Matches uppercase unicode alpha characters
- Replace Regular Expressions:
 - No characters
 - Preserve Source Value(s)

Buttons for OK, Cancel, and Close are at the bottom right.



The screenshot shows the main interface of Predictive Data Management v6.0.5. It includes a Repository Explorer on the left, a central workspace with a Data Map and Column Cyclone Operations, and an Actions panel on the right. The Column Cyclone Operations table is expanded to show the following data:

Group	Special Char.	Find	Re	Col By
Trim WhiteSpace	Valid Character			
Trim WhiteSpace	Valid Character			
Standardize	Valid Character			

Below the operations table is a Properties table for the [AMBDemoSource].[dbo].[Contact3].[LastName] data map:

ContactID	ContactID_status	LastName	LastName_status	LastName_clean	LastName_replace	LastName_group_i	LastName_match...
1374	NO ACTION	Salanki	MODIFIED	Salanki		3	Salanki
19982	NO ACTION	* + = % Smyth	MODIFIED	Smith	* + = % Smyth	1	Smith
19982	NO ACTION	Smyth	MODIFIED	Smith		1	Smith
6476	NO ACTION	Andersen	MODIFIED	Anderson		2	Anderson
780	NO ACTION	Smith-Bates	MODIFIED	Smith		1	Smith
784	NO ACTION	Solanki	MODIFIED	Salanki		3	Salanki
1483	NO ACTION	Solanki	MODIFIED	Salanki		3	Salanki
1486	NO ACTION	Smith-Bates	MODIFIED	Smith		1	Smith

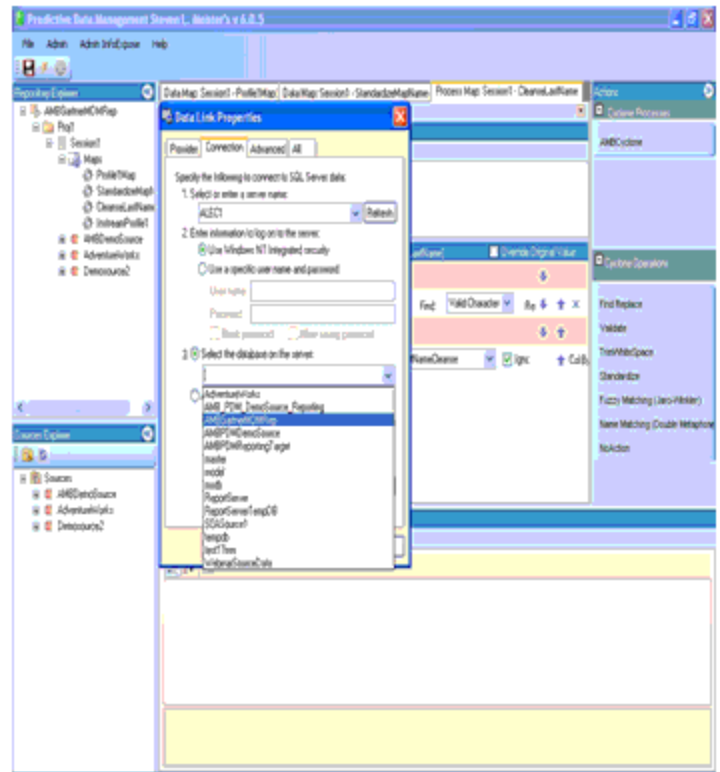
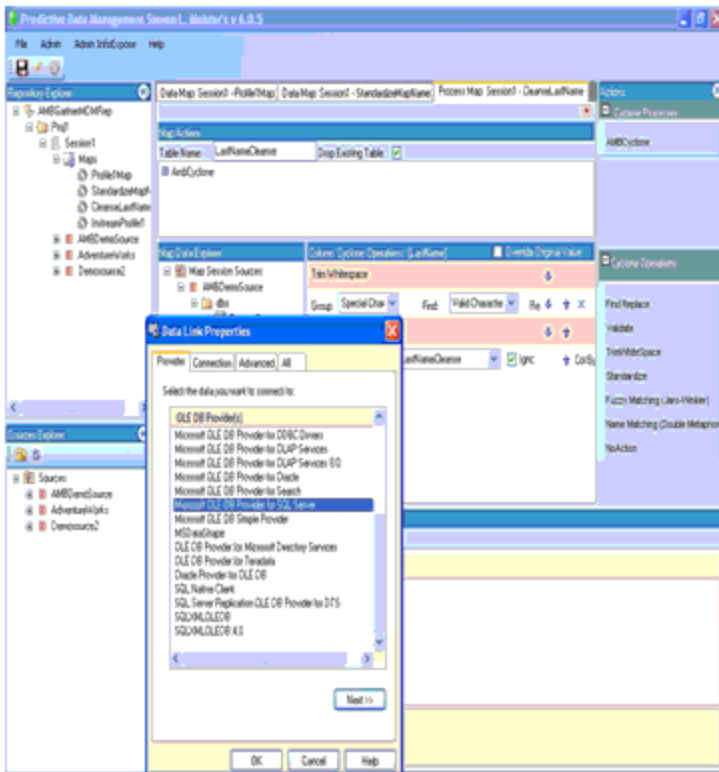
Other Significant Features

Sourcing Options:

PDM will source data directly from the originating data base. It does not require you to move data to a proprietary staging area. This saves considerable technicians time, computer resources and valuable project time to locate a place to move data to staging areas. PDM supports many environments for direct sourcing for profiling and quality, including DB2/UDB, DB2/400, DB2/z-OS Mainframe, Flat files, MS SQL, Oracle, and Teradata. PDM is one of the first supported products for SQL Server 2008 (as source, repository, and SSIS) with Version 6.0.5.

Destination Options:

PDM provides all the Profiling/Analysis results into an open repository rather than a vendor proprietary database. Repository database options include MS SQL Server, DB2/UDB and Oracle.



Address Verification:

PDM embeds CASS certified address verification for English speaking countries via our partner Satori Software.